# Evaluation of the Prediction Model and Oversample of Low Income Persons in the Medical Expenditure Panel Survey (MEPS)[1]

Lap-Ming Wun, Trena M. Ezzati-Rice, Robert Baskin,
Agency for Healthcare Research and Quality (AHRQ)
540 Gaither Road, Rockville, MD 20850

**Abstract**
The Medical Expenditure Panel Survey (MEPS) is sponsored by the Agency for Healthcare Research and Quality (AHRQ) and cosponsored by the National Center for Health Statistics (NCHS). It is conducted to provide nationally representative estimates of health care use, expenditures, sources of payment, and insurance coverage for the U.S. civilian noninstitutionalized population. It comprises three component surveys with the Household Component (HC) as the core survey. For selected years, the sample of the MEPS HC includes an oversample of several targeted subpopulations which include individuals with low income. However, the status of a household or dwelling unit's family income is not known at the time the MEPS sample is drawn. This characteristic has to be predicted. A predictive model for this characteristic was established using data from the 1987 National Medical Expenditure Survey (NMES), the predecessor of MEPS, and a 1986 screener survey. The model was evaluated and refined in 2001 and 2002 respectively using the latest MEPS data available then. This paper presents an evaluation of that model with more recent MEPS data, and an evaluation of the impact of the inclusion of oversampling by the predicted poverty status on variation of the MEPS weights. The result of this evaluation suggests that exclusion of oversampling by predicted poverty status may lead to improvement on population estimation on other subdomains.

**KEY WORDS:** logistic regression, predicted probability, accuracy, efficiency

## 1. Introduction

The Medical Expenditure Panel Survey Household Component (MEPS-HC) is an ongoing complex national probability survey of the civilian noninstitutionalized population and has been conducted since 1996 by the Agency for Healthcare Research and Quality (AHRQ). Data collected in the MEPS-HC provide nationally representative estimates of health care use, expenditures, sources of payment, and insurance coverage for the U.S. civilian noninstitutionalized population. The sampling frame for each year's MEPS HC sample is households participating in the previous year's National Health Interview Survey (NHIS) conducted by the National Center for Health Statistics of the Centers for Disease Control and Prevention. NHIS provides a nationally representative sample of the U.S. civilian noninstitutionalized population and oversamples Hispanics, blacks, and Asians (since 2006). Details of the MEPS sample design have been previously published (Ezzati-Rice et al, 2008).

In addition to the oversampling of Hispanics, blacks, and Asians inherited from the NHIS, MEPS also takes advantage of its unique linkage to the NHIS to oversample other policy relevant subgroups of interest. In particular, since 2002 the MEPS-HC has made a special effort to oversample the poor (those under 200% of the federal poverty level) using a model to predict those who will be poor a year later. Since the family income data from the NHIS are not available at the time of the selection of the MEPS sample, a prediction model is used to identify families predicted to have low income. Details of the development of the prediction model have been previously published (Moeller and Mathiowetz (1994)). The purpose of this paper is to present a summary evaluation of the stability and performance of the model over time, as well as the impact of the "predicted to be poor" oversampling on the variation of the MEPS weights.

## 2. The Prediction Model

A sample unit's income in a given year should be a reasonable predictor of its income status in the next year. However, studies as reported in Moeller and Mathiowetz (1994) have shown that the previous year's reported income on a screener interview is not a very reliable predictor for a subsequent survey year's poverty status due to under reporting and the dynamic of individuals moving into and out of poverty in adjacent years. Therefore, Moeller and Mathiowetz (1994)

[1]The views expressed in this paper are those of the authors and no official endorsement by the Department of Health and Human Services or the Agency for Healthcare Research and Quality is intended or should be inferred.

developed a predictive model based on the economic concept of permanent income, which is the family's expected income in a given year based on its human capital and other characteristics and resources. The model was estimated with data from the 1987 National Medical Expenditure Survey (NMES), the predecessor of the MEPS, and a screener interview conducted in 1986. A slightly modified version of the permanent income model identified the following variables as significant predictors of income status for MEPS sampling purposes:

1. Age of reference person.
2. Home ownership.
3. Reporting unit (RU) size.
4. Whether children of specific ages (under 6, 6-15) are present in the RU.
5. Whether someone in the RU other than the reference person is at least 65 years of age.
6. Health status of reference person.
7. Race/ethnicity of reference person.
8. Census Division.
9. Metropolitan statistical area (MSA) status and size of the primary sample unit (PSU).
10. Education of reference person.
11. Marital status and sex of reference person.
12. Whether reference person or spouse was employed in the previous 3 months.
13. Prior year's poverty status.
14. Whether anyone in the RU was covered by Medicaid.

Thirty-three bivariate variables were constructed for these 14 predictors. Using these variables as predictors and the poverty status classification as the dependent variable, a logistic regression model was developed to estimate the probability that a reporting unit would have a family income below the 200% poverty level in a subsequent year. Households with predicted probabilities above a certain threshold value were to be oversampled. Using the data from the 1987 NMES to examine the efficiency of various cut points as the threshold, it was determined that 0.3 was optimal in terms of the trade off between maximizing the sample yield and the accuracy of targeting the low income population. Consequently, all reporting units with a predicted probability of 0.3 or greater to have family income below the poverty level were oversampled.

The unit of analysis for the permanent income logistic regression model was the reporting unit (RU, i.e., family) (Cohen SB, 2000; Moeller and Mathiowetz, 1994). Estimates of the coefficients of the model were obtained using data from the 1987 NMES and the 1986 screener interview. After the equation was estimated, a year's NHIS data were used to calculate logit values. The logit value was then converted to a probability value for each NHIS RU, from which the next year's MEPS sample was to be drawn, through the equation:

PROB = EXP(LOGIT) / (1 + EXP(LOGIT));

This is the predicted probability that the sample unit would have family income below the poverty level in the next year. (Moore, 1997)

## 3. Evaluation of the Performance of the Model

Stability and performance of the model have been previously evaluated (Wun, L.M., Cohen SB, Moeller J, 2001 and 2002). The results showed that the coefficients of the model were relatively stable over time, and the model's performance was acceptable. However, additional elements have been recently added to the sample design of MEPS, e.g., additional oversampling by race/ethnicity. The additional minority oversampling in conjunction with the domain of predicted poverty status may have an impact on the resulting effective sample sizes. Therefore, we re-evaluate the performance of the model with more recent MEPS data. We then followup with an evaluation of the model's predictive effectiveness and the impact on variability of the MEPS weights.

### 3.1 Stability of the Coefficients

In a previous research evaluation the original estimated coefficients based on the 1987 NMES data were compared with those from the 1998 MEPS and the associated 1997 NHIS data. The results showed that out of 33 covariates, 3 had significant changes. In this current evaluation, we re-estimate the coefficients with more recent data, namely the 2004

MEPS and the associated 2003 NHIS, and we compare the estimated coefficients with the two earlier sets of coefficients. Based on the 2004 data, we observed that there were only 3 coefficients that had significant changes compared to those from 1998 and there were just 6 coefficients that had significant changes when compared to 1987 estimates. Thus, the coefficients are seen as relatively stable over time.

### 3.2 Accuracy and Efficiency of the Poverty Prediction Model

A second focus of this current research effort was to investigate the accuracy and efficiency of the prediction model. The sample selection of households for MEPS from the NHIS is generally carried out in the following manner. The NHIS responding households designated as eligible for MEPS subsampling are assigned to sample domains of analytic and policy level interest to MEPS. The overall sample size for each panel of the MEPS is determined according to the available budget and the eligible sample available from NHIS. The number of households selected from each sample domain depends on a variety of factors including available sample and analytic considerations. The household level sample domain variables vary slightly from year to year and include a hierarchical classification:

1. Any Asian in the household,
2. Any family in household predicted to be poor (below 200% of poverty level)
3. Any Hispanic in household
4. Any Black in household(not all years)
5. All others (i.e., no Asians, no families predicted as poor, no Hispanics, no Blacks)

Since 2002, the NHIS responding households eligible for MEPS that contained either any Asian or any family predicted to be poor were selected with certainty. For the 2005 MEPS used in this evaluation, blacks were selected at the rate of 0.75. For this evaluation of the accuracy and efficiency of the poverty prediction model, Asian households were excluded, since they were selected with certainty as the "poor" domain.

Table 1 shows the distribution of the 7,194 non-Asian households by predicted to be poor and actual poor based on the results from survey respondents in the 2005 panel 10 MEPS sample. For purposes of this paper, model accuracy is defined as the proportion of those dwelling units (DUs) predicted to be poor who actually turned out to be poor, i.e., the model's accuracy rate. Model efficiency is defined as among those DUs determined from the survey as actually poor, the proportion that was included in the sample based on the model predicted to be poor. For the 2005 MEPS (Panel 10), the model accuracy rate was approximately 58% and the model's efficiency rate was about 59%.

## 4. Evaluation of the Impact of the Oversampling by Poverty Status in the MEPS

### 4.1 Impact on Variance of Survey Weights

As with any survey which oversamples selected domains of the population, the differential sampling rates add to the variability of the weights for the population as a whole as well as for subdomains; thus, there is a reduction in the effective sample sizes. The design effect $(1+ CV^2)$ (Korn and Graubard, 1999) can be used to measure the impact of sample design on survey estimates resulting from the variation in the survey weights, table 2 shows that effect, especially for Hispanics and Blacks. For example: the last column in the table under Hispanic shows the numbers from 2005 MEPS panel 10. The means of the adjusted weights for the subgroups Predicted poor and Not predicted poor are 2,871 and 7,058 respectively. The respective coefficients of variation (CV) are 72.42% and 67.33%, and the design effects are 1.52 and 1.45. However, the CV of the entire group of Hispanics is 86.18% and the design effect is 1.74, which are much higher than those of either of the income subgroups. This is due to the difference between the two groups which is revealed by the difference between the means of the two subgroups. This between subgroup difference brought in the additional variation. This phenomenon exists in Hispanics and Blacks as shown in tables of these two groups. But it did not exist in Asians. Because Asian and poor households had the same sampling rate, the adjustment factors for poor and non-poor household in the group of Asians are the same. Therefore, there is no between subgroup difference induced by oversampling by poverty status among Asian households. The additional variation in the Other group is very small even though the adjustment factors between the Poor and Not poor subgroup are even larger than that of Hispanics or Blacks. This probably is because the proportion of poor households in this group is relatively small.

### 4.2 Impact on Sample Size if Predicted to be Poor were not used to Oversample

The oversampling by predicted to be poor in the MEPS induces additional variation to adjusted weights in other analytical domains and reduces the effective sample sizes for analytical purposes. To assess what would be the expected

number of households with low income in the MEPS sample if the predicted to be poor model was not used, we take advantage of the linkage between MEPS and NHIS to carry out such an evaluation. The expected sample size can be illustrated by the hypothetical example given below:

| | Current Sampling rate | | | Expected sampling rates |
|---|---|---|---|---|
| Predicted poor | 1.00 | | | 0.5 |
| Not predicted poor | 1/3 | | ALL | 0.5 |

In a sample frame of size 2000, of which 500 are predicted to be poor, the other 1500 are not. The sampling rate of the poor households is 1.00, and for the other group is 1/3, resulting in a sample of size 1000 of which 500 are poor, 500 are not. Now, eliminating the practice of oversampling, the expected sampling rate can be calculated from the sample size and the frame size. A sample of size 1000 from the frame of 2000 without oversamping of any subgroup, the sampling rate for everyone is 0.5 = (1000/2000). This is the expected sampling rate and resulted in a sample of 1000 in which 250 are poor. This 250 is the expected number of households predicted to be poor in the sample.

Based on this same rationale but with more complicated steps to account for the additional oversampling domains and the complex sample design of MEPS, we calculated the expected number of households in poverty for the MEPS years 2002 to 2005 if predicted to be poor was eliminated(from the 2001 to 2004 NHIS) as displayed in the 3rd row of Table 3. The calculation is based on redistributing the example year's sample size but without any oversampling by predicted poverty status. Row 1 of Table 3 also shows the actual number of households in poverty as determined from the MEPS sample via the household interview. We also estimated the effective sample size of households in poverty of the actual sample and approximated that of the expected sample if predicted poverty status is not used. The effective actual sample is the actual sample size taking into account the design effect associated with the variability of the sample weights (number of actual households divided by the factor $1+CV^2$ ). These effective sample sizes are shown in row 2 of Table 3. To approximate the effective expected sample size without the oversample by predicted poverty status we need to take out the effect of oversampling by predicted poverty status. This approximation is obtained by adjusting the weights of those households selected into the sample by their predicted poverty status to the weights that would have been had they were selected by race/ethnicity. For example, if a Black household in the 2005 MEPS (selected from 2004 NHIS) was selected due to its predicted poverty status, we divide its current actual weight by 0.75, the sample rate of Black households in 2005, to obtain the expected weight. This brings the weights of households to the level based on their race/ethnicity status and eliminates the effect of oversampling by predicted poverty status. Using the adjusted weights, we approximated CV's for the year, and use these CV's to calculate the effective expected sample size without the oversample by predicted poverty status, the resulting effective sample sizes are given in the last row of table 3. From table 3, the expected number of households in poverty (row 3) is smaller than the actual number of households with poverty oversampling (row 1), however the reduction in effective sample sizes (row 2 to row 4) is much smaller.

## 5. Discussion

The MEPS sample design reflects multiple analytical goals. Oversampling is a key feature of the MEPS sample design helping to increase the sample sizes and thus improve the precision of estimates for selected subgroups of the population to enhance policy relevant analysis. The oversampling, however, adds to the variability of the MEPS survey weights and in some cases reduces the effective sample sizes. Since 2002, MEPS has targeted the "poor" for oversampling using a model that takes advantage of the MEPS and NHIS survey linkage. The model attempts to predict those who will be poor in a year's time and in two years. MEPS also oversamples selected race/ethnic groups. In this paper, we have shown that the coefficients used in the prediction model have remained very stable over time. However, when we examined the accuracy and efficiency of the model, the rates were determined to only be about 60 percent. While not shown in this paper, the accuracy and efficiency rates at the person level were even lower. In addition, the intersection of the oversampling for those under 200 percent of the poverty level with the oversampling by race/ethnicity, contributes to the variability in the weights for the population overall. Exclusion of the "predicted to be poor" as an oversampling domain can reduce the variability in the weights**,** in particular for the full population and specific domains, e.g., Blacks and Hispanics. This would likely outweigh the effect of the small reduction in the effective number of poor households.

## References

Cohen SB. Sample Design of the 1997 Medical Expenditure Panel Survey Household Component. Rockville (MD): Agency for Healthcare Research and Quality; 2000. MEPS Methodology Report No. 11. AHRQ Pub. No. 01-0001.

Ezzati-Rice, TM, Rohde F, Greenblatt, J. Sample Design of the Medical Expenditure Panel Survey Household Component, 1998-2007. Methodology Report No. 22. March 2008. Agency for Healthcare Research and Quality, Rockville, MD. http://www.meps.ahrq.mepsweb/data_files/publications/mr22/mr22.pdf

Korn EL, Graubard BI (1999) Analysis of Health Surveys, John Wiley & Sons, New York.
Moeller J, Mathiowetz N. "Problems of Screening for Poverty Status", Journal of Official Statistics, 1994, Vol. 10, No. 1, pp. 327-337.

Moore G. Identification of the MEPS 1997 sample from the 1996 NHIS. Bethesda (MD): Social and Scientific Systems, Inc.; 1997. Task NMS2.112 Report.

Wun, L.M., Cohen SB, Moeller J (2001), "An Evaluation of a Model used to Oversample Low Income Households in the 1997 MEPS," 2001 Proceedings of the American Statistical Association, Survey Research Method Section (CD-ROM), Alexandra, VA American Statistical Association.

Wun, L.M., Cohen SB, Moeller J (2002), "Refining the Prediction Models of Low Income Status and Future Medical Expenditures in the Medical Expenditure Panel Survey ," 2002 Proceedings of the American Statistical Association, Survey Research Method Section (CD-ROM), Alexandra, VA American Statistical Association.

**Table 1. Model accuracy and efficiency using 7,194 Dwelling Units (DU) in the 2005 MEPS (Panel 10) sample that are non- Asian**:

| | | Predicted poor | | |
|---|---|---|---|---|
| | | No | Yes | |
| Actual poor | No | 4257 | 892 | 5149 |
| | Yes | 831 | 1214 | 2045 |
| | | 5088 | 2106 | 7194 |

Model accuracy: 1,214/2,106 ≈ 58%
Model efficiency: 1,214/2,045 ≈ 59%

**Table 2. Coefficients of Variation (CVs), design effects (deffs), and Means (in ( )) of Weights, one panel, first year, by race/ethnicity.**
**In each cell: the first number is CV, the second number is deff, number in ( ) is the mean.**

Hispanic

| | 2002 P7 | 2003 P8 | 2004 P9 | 2005 P10 |
|---|---|---|---|---|
| | | | | |
| All | 92.09, 1.85, (3914) | 75.82, 1.57, (4710) | 90.07, 1.81, (4536) | 86.18, 1.74 , (4881) |
| | | | | |
| Not predicted poor | 69.04, 1.48, (6233) | 62.54, 1.39, (6347) | 73.94, 1.55, (6562) | 67.33, 1.45 , (7058) |
| Predicted poor | 68.52, 1.47, (2235) | 59.97, 1.36, (2865) | 64.86, 1.42, (2639) | 72.42, 1.52 , (2871) |

Black

| | 2002 P7 | 2003 P8 | 2004 P9 | 2005 P10 |
|---|---|---|---|---|
| | | | | |
| All | 77.07, 1.59, (6045) | 66.50, 1.44, (6908) | 65.62, 1.43, (7251) | 61.24, 1.37, (6183) |
| | | | | |
| Not predicted poor | 57.58, 1.33, (8625) | 48.89, 1.24, (9540) | 56.30, 1.32, (9049) | 54.28, 1.29, (7361) |
| Predicted poor | 56.98, 1.32, (3224) | 50.86, 1.26, (3981) | 54.06, 1.29, (4618) | 55.33, 1.31, (4217) |

Asian

| | 2002 P7 | 2003 P8 | 2004 P9 | 2005 P10 |
|---|---|---|---|---|
| | | | | |
| All | 49.85, 1.25, (8125) | 52.35, 1.27, (8533) | 56.29, 1.32, (9967) | 52.96, 1.28, (9908) |
| | | | | |
| Not predicted poor | 48.35, 1.23, (8469) | 50.49, 1.25, (8784) | 55.21, 1.30, (10014) | 52.90, 1.28, (10128) |
| Predicted poor | 52.44, 1.27, (6484) | 63.70, 1.40, (6651) | 67.05, 1.44, (9506) | 49.54, 1.24, (8430) |

Other

| | 2002 P7 | 2003 P8 | 2004 P9 | 2005 P10 |
|---|---|---|---|---|
| | | | | |
| All | 50.94, 1.26, (10422) | 48.49, 1.24, (11537) | 52.29, 1.27, (11742) | 50.85, 1.26, (12137) |
| | | | | |
| Not predicted poor | 43.26, 1.19, (11580) | 42.18, 1.18, (12630) | 45.75, 1.21, (12866) | 44.00, 1.19, (13297) |
| Predicted poor | 45.09, 1.20, (4736) | 40.86, 1.17, (5572) | 46.16, 1.21, (5458) | 59.79, 1.36, (6083) |

**Table 3. Actual vs. expected number of households in poverty in the 2002-2005 MEPS samples**

| row | | From NHIS 2001 | From NHIS 2002 | From NHIS 2003 | From NHIS 2004 |
|---|---|---|---|---|---|
| 1 | Actual | 1982 | 1791 | 1867 | 2094 |
| 2 | Actual (effective)* | 1318 | 1259 | 1253 | 1419 |
| 3 | Expected | 1495 | 1464 | 1532 | 1765 |
| 4 | Expected(effective)* | 1234 | 1244 | 1274 | 1367 |

* Effective sample size = sample size / $(1+CV^2)$

Where CVs of means of the Actual weights for each of the years 2001, 2002, 2003, and 2004 are: 0.71, 0.65, 0.70, and 0.69, respectively, those of the Expected samples are:0.46, 0.42, 0.45, and 0.54.